

Simple Models of Genomic Variation in Human SNP Density

Raazesh Sainudiin*, Andrew G. Clark†, and Richard T. Durrett‡

*Department of Statistical Science, Cornell University, Ithaca, New York 14853.

†Department of Mol. Biology and Genetics, Cornell University, Ithaca, New York 14853

‡Department of Mathematics, Cornell University, Ithaca, New York 14853.

Summary. Hierarchical Poisson models and coalescent mixture models are used to explain the observed variation in single-nucleotide polymorphism (SNP) density across the human genome. Accounting for heterogeneities in mutation as well as the recombination rates can better fit the observed human SNP density distribution.

0.1. Introduction

Understanding the population genetic forces behind the observed variation among human genome sequences is vital to deciphering the genetic causes of phenotypic variation among human beings. Suppose that the genomes of two randomly sampled human beings were tracked back through time, then one would observe several biological phenomena that shape the two genomes until the entirety of the two genomes have coalesced into one. These phenomena include variation-introducing events, such as, point mutations, recombinations, and activities of various transposable elements, as well as, coalescent-affecting scenarios, such as, population dynamics, population structure, and natural selection. A mechanistic understanding, by means of explicit population genetic models of coalescence in the presence of recombination and mutation, of the observed genomic variation in SNP density must address any interplay among the heterogeneities in the above phenomena. This study is a first order approximation to this challenging problem. We only focus on the interplay between the heterogeneities in point mutation rate and recombination rate to explain the observed human SNP density as determined with the Celera-PFP SNPs shown in Figure 14 of the study by Venter et al. (2001) that examined genome-wide sequence variation. Our SNP density data was obtained from first aligning the Celera consensus sequence to the PFP assembly and then counting the number of SNPs in bins of 100 kbp (100,000 basepairs), as described in section 6 of Venter et al. (2001). Thus, we are only interested in random samples of size 2 from a locus that is 100 kbp in length.

Two approaches toward modelling are taken. The first approach, by means of hierarchical Poisson models, obtains better fits than the homogeneous Poisson distribution used by Venter et al. (2001). Insights gained from the first approach inform the second approach. The second approach is more mechanistic as it employs mixtures of SNP densities simulated under the coalescent with different mutation and recombination rates to obtain a better fit to the observed SNP density distribution. This approach introduces heterogeneity into the coalescent-based simulation of SNP density that was shown to produce a poor fit under the assumptions of genome-wide homogeneity and equality of mutation and recombination rates (Venter et al., 2001).

0.1.1. Hierarchical Poisson Models:

Let Λ and T be the parameters in the mass function of a Poisson distribution given by $\Pr(X = x|\Lambda T) = e^{-\Lambda T} (\Lambda T)^x / x!$. The relative mutation rate is represented by Λ . The sum of branch lengths

Table 1. Maximum likelihood analysis of Poisson models.

Model	T	Λ	Maximum Likelihood Estimates	ML
Poisson	λ	1	$\hat{\lambda} = 90.2, \hat{\mu} = 90.2, \hat{\sigma}^2 = 90.2$	-616497
A	$G(\gamma_1, \gamma_2)$	1	$\hat{\gamma}_1 = 2.7, \hat{\gamma}_2 = 32.9, \hat{\mu} = 90.2, \hat{\sigma}^2 = 3049.7$	-186348
A'	λ	$B(\beta_1, \beta_2)$	$\hat{\lambda} = 387.6, \hat{\beta}_1 = 2.17, \hat{\beta}_2 = 7.16, \hat{\mu} = 90.1, \hat{\sigma}^2 = 2683.9$	-185869
B	$G(\gamma_1, \gamma_2)$	$B(\beta_1, \beta_2)$	$\hat{\gamma}_1 = 6.4, \hat{\gamma}_2 = 19.0, \hat{\beta}_1 = 1.3, \hat{\beta}_2 = 0.46, \hat{\mu} = 90.1, \hat{\sigma}^2 = 2538.2$	-185511

of the coalescent trees for all the non-recombining segment(s) of the 100kbp locus is represented by T . In other words, T is the sum of the branch lengths of the ancestral recombination graph of our sample of size 2 at a locus that is 100kbp long. The random variable X represents the count of SNPs in contiguous 100kbp intervals from an alignment of two human genomes. In this hierarchical scheme, heterogeneities are modeled by the following Gamma and Beta probability density functions (PDFs),

$$T \sim G(\gamma_1, \gamma_2) \Leftrightarrow PDF(t) = \frac{1}{\Gamma(\gamma_1) \gamma_2^{\gamma_1}} t^{\gamma_1-1} \exp\left(-\frac{t}{\gamma_2}\right), \quad 0 \leq t < \infty, \gamma_1, \gamma_2 > 0,$$

$$\Lambda \sim B(\beta_1, \beta_2) \Leftrightarrow PDF(\lambda) = \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)} \lambda^{\beta_1-1} (1-\lambda)^{\beta_2-1}, \quad 0 \leq \lambda \leq 1, \beta_1, \beta_2 > 0.$$

In the hierarchical Poisson model A, we allow $T \sim G(\gamma_1, \gamma_2)$, while Λ is fixed at 1. The fit to the data (Figure 1) improved in comparison to the homogeneous Poisson fit which completely ignores the underlying ancestral recombination process. Thus, when the Gamma distribution is used to approximate the distribution of the sum of all branch lengths of the ancestral recombination graph of a locus whose recombination rate was drawn from some unknown distribution, the observed variance is better explained.

In model A, the mutation rate parametrized by Λ was fixed. In order to allow variation in mutation rate, we let $\Lambda \sim B(\beta_1, \beta_2)$. A hierarchical Poisson model A' that restricts T to a constant parameter λ while allowing Λ to be Beta distributed was fit to the data. The fit was better than that of model A. Thus, modelling heterogeneity in mutation rates across the different 100kbp loci gives better fits to the SNP density distribution. When we allowed both Λ to be Beta distributed and T to be Gamma distributed, we get the hierarchical Poisson model B.

As shown in Figure 1, the fit is much better to the observed data when heterogeneities in both mutation and recombination are accounted for through model B. The results of the maximum likelihood (ML) analysis of these four Poisson models are summarized in Table 0.1.1. The first and second moments ($\hat{\mu}$, and $\hat{\sigma}^2$) under the maximum likelihood estimates are also shown for each model in the Table. Note that the means are almost the same but the variances vary considerably. Guided by insights from these phenomenological hierarchical Poisson models, we analyze mechanistic models of the neutral coalescent with heterogeneities in both mutation and recombination rates under a simulated maximum likelihood framework.

0.1.2. Coalescent Mixtures:

A panmictic, Wright-Fisher, neutral coalescent model with a constant effective population size of 10,000 diploid individuals was assumed to simulate the distribution of the number of segregating sites at a locus of 100 kbp evolving under an infinite sites mutation model using the C program *ms*

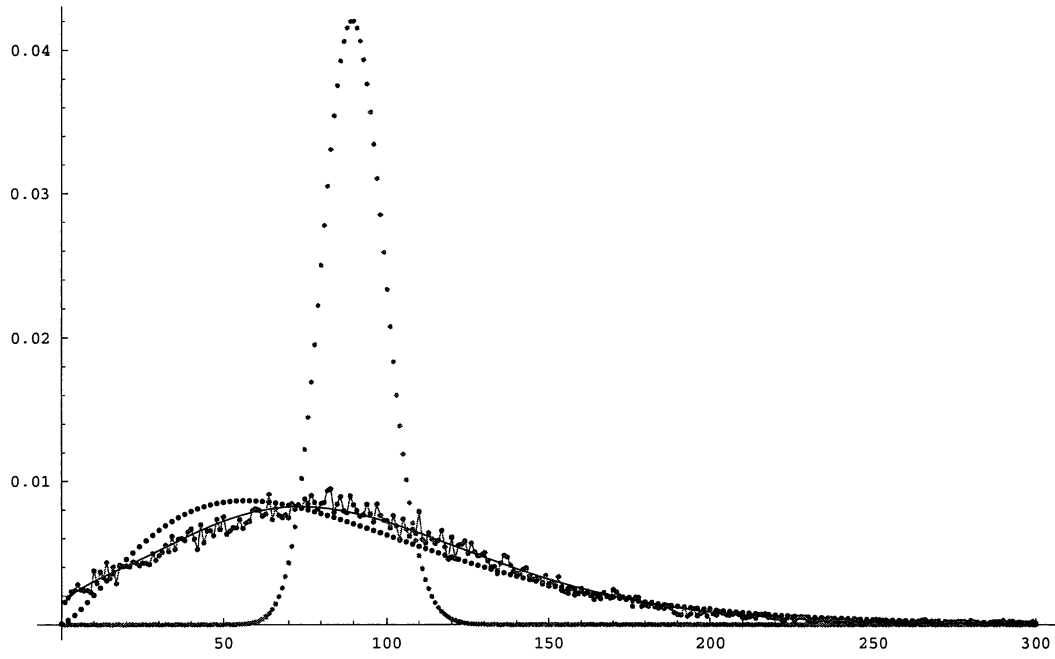


Fig. 1. Fits of the homogeneous Poisson model (large gray dots), hierarchical Poisson model A (black dots) with $T \sim G(\gamma_1, \gamma_2)$, and hierarchical Poisson model B (black line) with $T \sim G(\gamma_1, \gamma_2)$ and $\Lambda \sim B(\beta_1, \beta_2)$ to the observed SNP density distribution (joined gray dots).

(Hudson, 2002). The scaled product of the effective population size (N_e) and the mutation rate per locus per generation (μ) is denoted by $\theta = 4N_e\mu$. The recombination rate r is the probability of cross-over per generation between the ends of the locus being simulated and its scaled product with N_e is denoted by $\rho = 4N_er$.

In the complete absence of recombination, the distribution of SNPs under the above assumptions have an explicit form. The coalescent tree is identical for every nucleotide site in the locus in any given realization of the coalescent process of two samples. Since the rescaled time to the coalescent event and the mutation event are exponentially distributed with rates 1 and θ , respectively, the probability of a mutation event before the coalescent event is $\theta/(1 + \theta)$. Thus, the probability of observing x mutations at our locus before the coalescent event is $(\theta/(1 + \theta))^x 1/(1 + \theta)$. In other words, the probability of observing x SNPs at a locus with $r = 0$ is geometrically distributed with parameter $1/(1 + \theta)$.

If the recombination rate at our locus approaches infinity, then the distribution of SNPs approaches a Poisson distribution with parameter θ . This can be seen from the following argument. High levels of recombination assures that the coalescent tree at each site is independent of those at other sites. Thus, for a locus with n sites, the probability of the observing x SNPs is $\binom{n}{x} (\frac{\theta}{n}/(1 + \frac{\theta}{n}))^x (1/(1 + \frac{\theta}{n}))^{n-x}$. For large loci, this binomial mass function is known to approximate $e^{-\theta}\theta^x/x!$, the Poisson mass function, as $n \rightarrow \infty$ and $n \frac{\theta}{n}/(1 + \frac{\theta}{n}) \rightarrow \theta$.

However, when the recombination rate is some intermediate value between the above two extremes no explicit forms are known for the SNP density. We use empirical estimates of the SNP density from a large number of simulations (typically 100,000). Figure 2 shows how the distribution of SNP density under our assumptions morphs from the geometric distribution (black dots) towards the Poisson distribution (gray dots) as the scaled recombination rate ρ increases from 0 to 1000 in decreasing shades of gray. This behavior is identical for any fixed value of θ except for a scale change.

The empirical estimates of the sex-averaged human recombination rates in 1 Mbp intervals based on Genethon map were downloaded from the UCSC annotation database (UCSC, UCSC). We interpolated to obtain the estimates over 100 kbp segments by assuming rate constancy over the 10 consecutive 100 kbp segments that constitute the 1 Mbp segment for which an empirical estimate of the recombination rate were available. Let this empirical distribution of the sex-averaged human recombination rate in 100 kbp intervals based on Genethon map, as shown in Figure 3, be denoted by \hat{R} . The following strategy was used to obtain an estimate of the SNP density distribution for each scaled mutation rate $\theta_i \in \Theta = \{\theta_1, \dots, \theta_{304}\} = \{0.001, 0.01, 0.1, 0.5, 1, 2, \dots, 300\}$, when the recombination rate was assumed to be distributed according to \hat{R} .

- for each $\theta_i \in \Theta$, repeat N times:
 - sample a ρ according to \hat{R}
 - simulate the described coalescent with θ_i
 - record the number of SNPs
- Obtain the empirical distribution of SNP density for the given θ_i when $\rho \sim \hat{R}$

We denote this simulation-based estimate of the SNP density distribution for each $\theta_i \in \Theta$ by $\hat{S}_{\hat{R}, \theta_i}$. Note that $\hat{S}_{\hat{R}, \theta_i} \rightarrow S_{\hat{R}, \theta_i}$, the true SNP density distribution, as the number of replicates (N) used to estimate it grows large. In practice, N was set at 100,000 since an increase in N to 150,000 replicates did not significantly affect the estimates. A discretized and rescaled Beta density with parameters α and β was used to find the mixing weights for each $\theta_i \in \Theta$. Thus, for every pair (α, β) ,

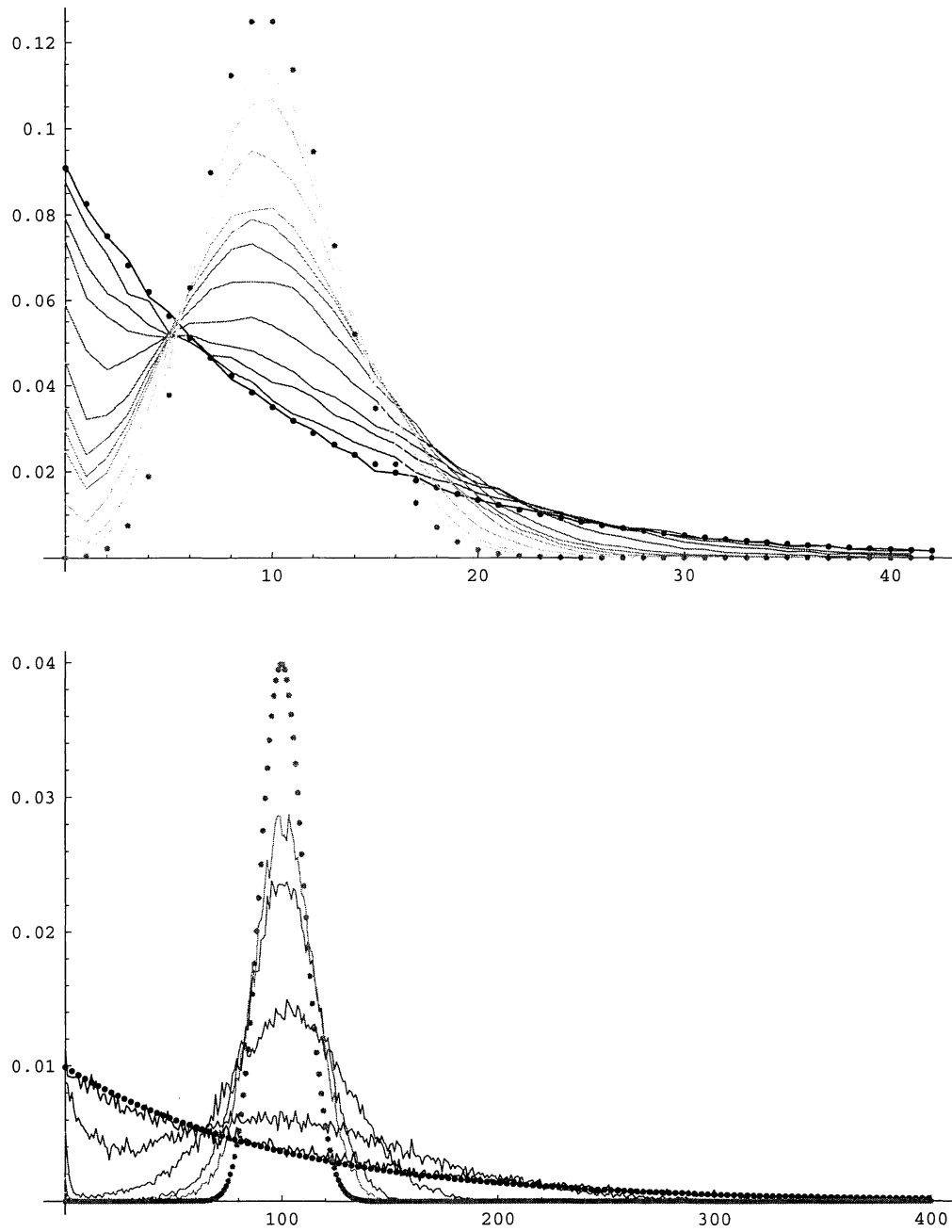


Fig. 2. The distribution of SNP density in 100 kbp morphs from the geometric distribution (black dots) towards the Poisson distribution (gray dots) as the scaled recombination rate ρ increases from 0 to 1000 in decreasing shades of gray for $\theta = 10$ (top) and $\theta = 100$ (bottom).

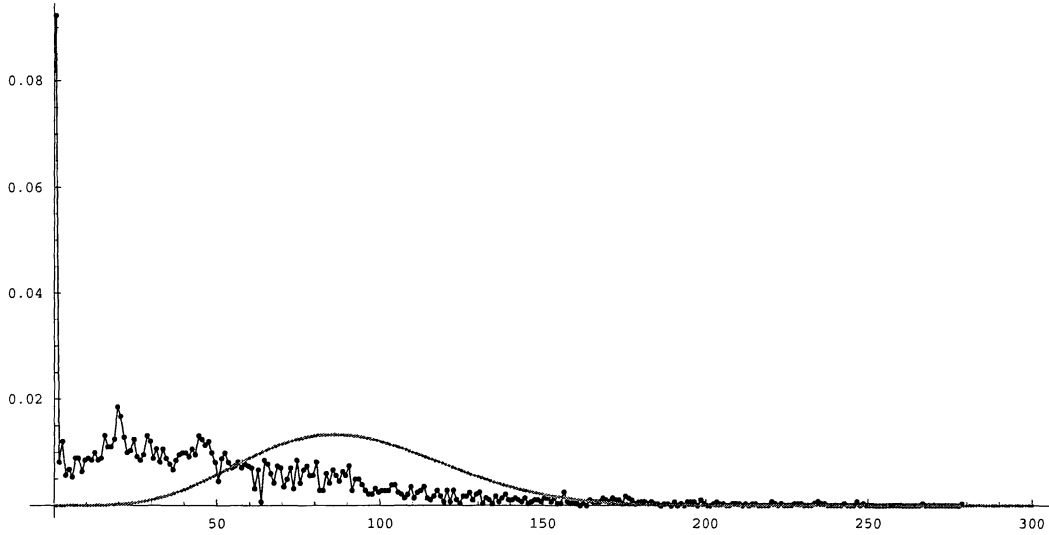


Fig. 3. The distribution of the empirical estimates of the sex-averaged recombination rate in 100 kbp segments of the human genome from the Genethon map (joined black dots) and $w_{\theta_i}^{(6.7, 14.9)}$, the maximum simulated likelihood estimate of the weights on $\theta_i \in \Theta$ (gray line) for the coalescent mixture model.

the shape of the Beta density specified the mixing weights, as follows:

$$w_{\theta_i}^{(\alpha, \beta)} = \int_{\frac{i-1}{|\Theta|}}^{\frac{i}{|\Theta|}} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \lambda^{\alpha-1} (1 - \lambda)^{\beta-1} d\lambda$$

where, $|\Theta| = 304$ is the cardinality of the set Θ . Such (α, β) -specified $w_{\theta_i}^{(\alpha, \beta)}$'s were used to weigh the corresponding $\hat{S}_{\hat{R}, \theta_i}$'s in order to obtain a finite mixture of the form $\sum_{\theta_i \in \Theta} w_{\theta_i}^{(\alpha, \beta)} \hat{S}_{\hat{R}, \theta_i}$. A simulated likelihood function of α and β was thus constructed for the given SNP data $X = (x_0, \dots, x_n)$, as follows,

$$\prod_{j=0}^n \sum_{\theta_i \in \Theta} w_{\theta_i}^{(\alpha, \beta)} \hat{S}_{\hat{R}, \theta_i}(x_j)$$

Local Quasi-Newton searches were performed to find the maximum simulated likelihood (*MSL*) estimates $\hat{\alpha} = 6.7$ and $\hat{\beta} = 14.9$ ($MSL = -185555$). We also did a least-squares fit of the observed to the predicted densities and found comparable estimates. Empirical estimates of the sex-averaged recombination rates from deCODE, and Marshfield maps were also used in a similar analysis. Comparable estimates were obtained under a reasonably good fit ($MSL = -185558$) with the deCODE map whose empirical CDF resembles that of the Genethon Map. However, an analysis with the Marshfield map yielded a poorer fit ($MSL = -186007$). Figure 4 summarizes the fits to the observed SNP data while Figure 3 shows the marginal density of ρ from the Genethon map and the marginal density of θ under the maximum simulated likelihood estimates ($\hat{\alpha} = 6.7$, $\hat{\beta} = 14.9$) with mean and variance given by 90.7 and 876.1, respectively.

Another study by international SNP map working group (2001) claimed to have achieved a good fit to single reads with 0, 1, 2, 3, or 4 SNPs, by accounting for mutational heterogeneity

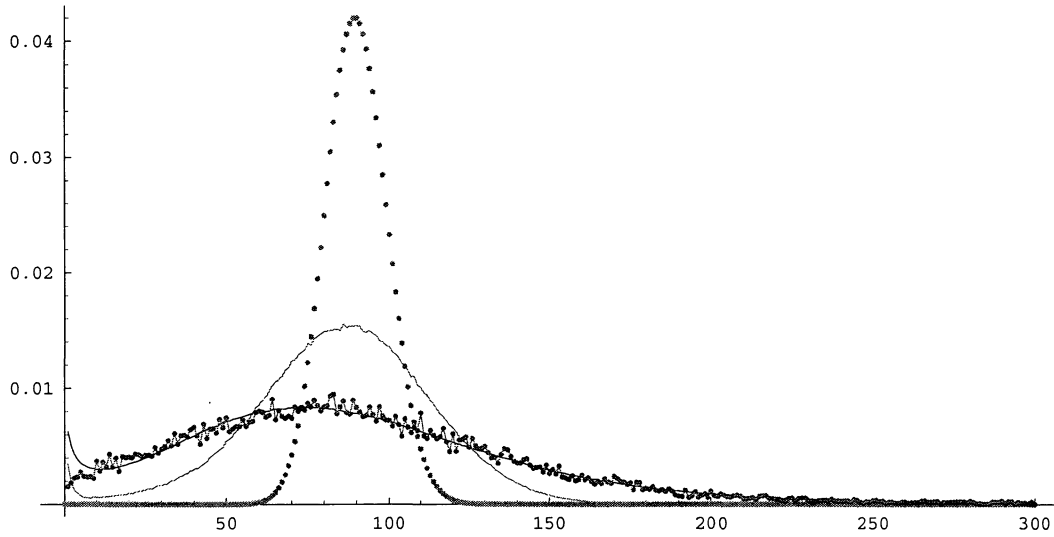


Fig. 4. The SNP density distribution (joined gray dots), Poisson distribution with mean 90 (large gray dots), Simulated distribution of SNPs with $\rho = \theta = 90$ (gray line), and the Maximum simulated likelihood estimate from the coalescent simulations with $\rho \sim \hat{R}$ and $\theta_i \sim w_{\theta_i}^{(6.7, 14.9)}$ (black line).

and genealogical variability in a different manner. They partitioned the genome into 200 kbp bins, and selected a read from each bin. They calculated the observed GC content of the bin, and from a regression of GC content on nucleotide diversity across the whole genome, they calculated an expected diversity given the local GC content of each bin induced by the exponentially distributed coalescent time for samples of size 2 in the absence of recombination. However, Venter et al. (2001) used the full bin size of 100 kb, so the SNP count ranged to more than 100 per bin. Because many neighboring reads have shared genealogies, the magnitude of variability from bin to bin is much greater, and the power to detect this heterogeneity is far greater. Thus, Venter et al. (2001) found that the coalescent in the presence of recombination fit the observed SNP density better than the coalescent without recombination. The model of international SNP map working group (2001) fits without recombination only because the power is so low to detect a departure. Using the data of Venter et al. (2001) in this study, we find that coalescent with heterogeneities in recombination as well as mutation gives substantially better fits than coalescent with a constant rate of recombination and mutation.

We have shown that by invoking heterogeneities in mutation and recombination rates, one can better explain the observed variation in SNP density across 100 kbp segments of two randomly sampled human genomes. Phenomenological fits by means of hierarchical Poisson models, as well as mechanistic fits by means of coalescent mixture models, significantly improved when heterogeneities in recombination as well as mutation rates were accounted for. The coalescent mixture model does not completely fit the data in the most interesting region, namely, the segments with the least SNP density. This is partly due to the filtering strategy used to obtain the data. Since there were considerable gaps in the alignments for several bins, there was an overestimation of bins with 0 SNPs. Thus, these bins were ignored from the analysis by Venter et al. (2001). As high resolution data for larger samples become available at a genomic scale, one can use more sophisticated simulated

ML methods to get the null distributions of various test statistics under the Wright-Fisher neutral coalescent that account for empirically observed heterogeneities in recombination and mutation rates. Heterogeneities in well-understood and observed biological phenomena, such as recombination hot spots, should be used in the null hypothesis when more complex models with unobserved phenomena, such as population dynamics, population structure, and/or natural selection are tested.

1. ACKNOWLEDGMENTS

R.S. and R.D. are partially supported by the National Science Foundation/National Institutes of Health Grant DMS/NIGMS 0201037 to Durrett, Aquadro, and Nielsen. R.S. would like to thank Arkendra De, Kevin Thornton, and Russell Zaretzki for insightful discussions.

References

- Hudson, R. R. (2002). Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics* 18, 337–338.
- international SNP map working group, T. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928–933.
- UCSC. <http://genome.ucsc.edu/goldenPath/gbdDescriptions.html>.
- Venter, C. J., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, and *et al.* (2001). The sequence of the human genome. *Science* 291, 1304–1351.